Gutiérrez Muñoz, Francisco
Rey Guerrero, Alfredo Del

ESTUDIO ESTADISTICO DE PALABRAS Y CARACTERES EN TITULOS DE ARTICULOS CIENTIFICO–TECNICOS EN ESPAÑOL

Resumen

Se muestran los resultados de un estudio estadístico efectuado a partir de una colección de títulos de artículos científico-técnicos en español, procedentes de la base de datos del Indice Español de Ciencia y Tecnología. En dicho estudio se han determinado las longitudes de los títulos (número de palabras por título) y de las palabras (número de caracteres por palabra), y se han relacionado con las frecuencias de aparición de las palabras en los títulos y de las letras en las palabras. Se ha examinado la incidencia de las palabras poco significativas (vacías) y se muestra una lista de las palabras vacías que aparecen con mayor frecuencia en los títulos de la colección.

Palabras clave: *Títulos de artículos. Estudio estadístico. Análisis de frecuencias. Frecuencia de palabras. Frecuencia de letras. Palabras vacías. Español.*

Abstracts

The results of a statistical study carried out on a series of titles of scientific and technical documents, from the secondary publication Indice Español de Ciencia y Tecnología, are exposed. In this study have been calculated the number of words by title and the number of characters by word. The have been related with the appearing fequequency of words in titles and of letters in words. The incidence of the stop word was examined and a table is give with the more frequent stopwords appearing in the titles.

Keywords: *Document Titles. Statistical Study. Frequency-Analysis. Word Frecuency. Letter. Frequency. Stopwords. Spanish.*

Introducción

El Indice Español de Ciencia y Tecnología (IECYT) es una publicación periódica editada por el Instituto de Información y Documentación en Ciencia y Tecnología del CSIC. que da cuenta de los trabajos que aparecen publicados en más de 200 revistas científicas y técnicas españolas, lo cual le permite referenciar la gran mayoría de los artículos que se publican en España en relación con cualquier campo de la ciencia y la tecnología, a excepción de la medicina.

En este trabajo se muestran los resultados obtenidos en relación con un recuento estadístico de palabras y caracteres que se ha efectuado utilizando una colección de 12.540 títulos de artículos científicos y técnicos en español, obtenidos a partir de la base de datos del IECYT. Los títulos incluidos en la colección se corresponden estrictamente con la información en español recogida entre 1980 y 1984 por el IECYT. Por ello, esta colección proporciona una muestra muy rica y variada de la terminología utilizada en la elaboración de los títulos de los artículos, y con un reflejo de los temas que interesan a un colectivo numeroso de especialistas de muy diversos campos. En la Figura 1 se muestra gráficamente la distribución temática de los títulos incluidos en la colección.

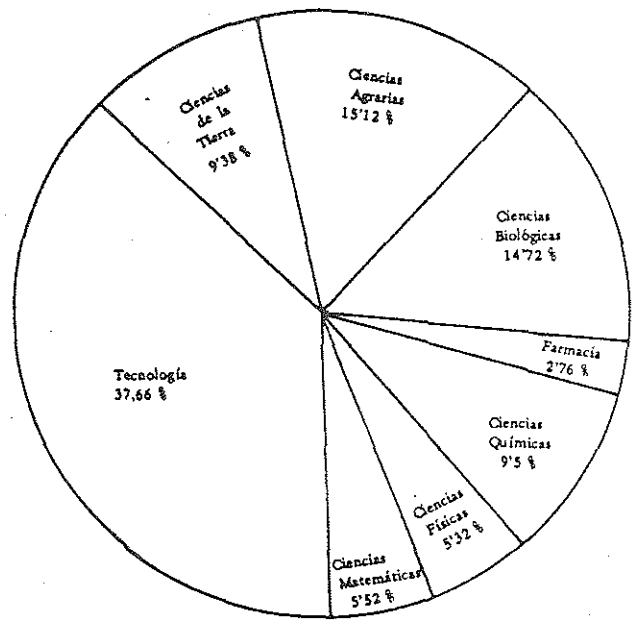


Figura nº 1

Metodología

Para llevar a cabo el estudio estadístico que nos ocupa ha sido necesario desarrollar una serie de programas en lenguaje COBOL, que partiendo de la base de datos del IECYT, han permitido:

– Seleccionar los títulos de artículos en español que integran la colección con la que se ha trabajado.

– Independizar las palabras contenidas en dichos títulos.

– Clasificar alfabéticamente las palabras independizadas y determinar su frecuencia de aparición.

– Determinar posiciones y frecuencias de aparición de letras en palabras.

– Relacionar estadísticamente títulos, palabras y letras.

Sobre la base de los listados obtenidos se llevó a cabo posteriormente, un examen de las palabras extraídas, con el fin de determinar porcentajes de palabras "vacías" y de palabras con contenido.

Antes de exponer los resultados obtenidos conviene señalar que se ha preferido trabajar con títulos originales en español, con el fin de que la terminología de trabajo fuera exclusivamente la utilizada por los propios autores.

Asimismo, conviene advertir que en este trabajo se ha considerado que una palabra está constituida por cualquier secuencia de caracteres alfanuméricos que comience por una letra, y que esté delimitado por caracteres especiales de cualquier tipo (tales como signos de puntuación, espacios en blanco, etc.).

Por consiguiente, se han considerado como palabras términos tales como: 1) sustantivos, adjetivos, verbos, adverbios, artículos, pronombres, conjunciones y preposiciones; 2) abreviaturas y acrónimos; 3) términos grecolatinos o procedentes de otros idiomas incluidos en los títulos; 4) nombres propios; 5) letras sueltas; 6) números romanos, y 7) fórmulas químicas.

En este trabajo, y en relación con las palabras, también se ha convenido lo siguiente : palabras homógrafas u homónimas se consideran como idénticas; p. ej.: *vino* (bebida) y *vino* (del verbo venir). También se consideran como iguales las palabras que sólo se diferencian por su acentuación; p. ej.: *hallo* y *halló*. Las palabras que contengan errores ortográficos o tipográficos son consideradas como distintas en relación con sus versiones correctas; p. ej.: *huevo* y *huebo*. Cuando varias palabras aparecen yuxtapuestas, ya sea por error o intencionadamente, el término resultante es considerado como una sola palabra; p.: ej. *fisicoquímica, métododе,* etc. En cambio, se consideran como palabras diferentes las que aparecen relacionadas mediante guiones; p. ej.: *físico-química* . En una expresión aritmética se consideran como palabras sólo a los miembros u operandos que comiencen por una letra.

Títulos

De los 12.540 títulos incluidos en la colección se extrajeron en total 142.944 palabras. El número de palabras extraídas por título ha oscilado entre 1 y 48, si bien el promedio ha sido de 11,3 palabras por título.

Figura nº 2

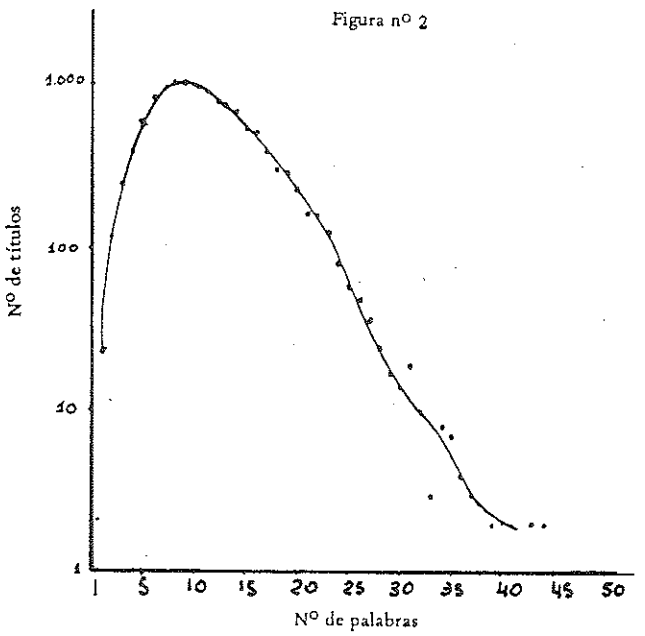
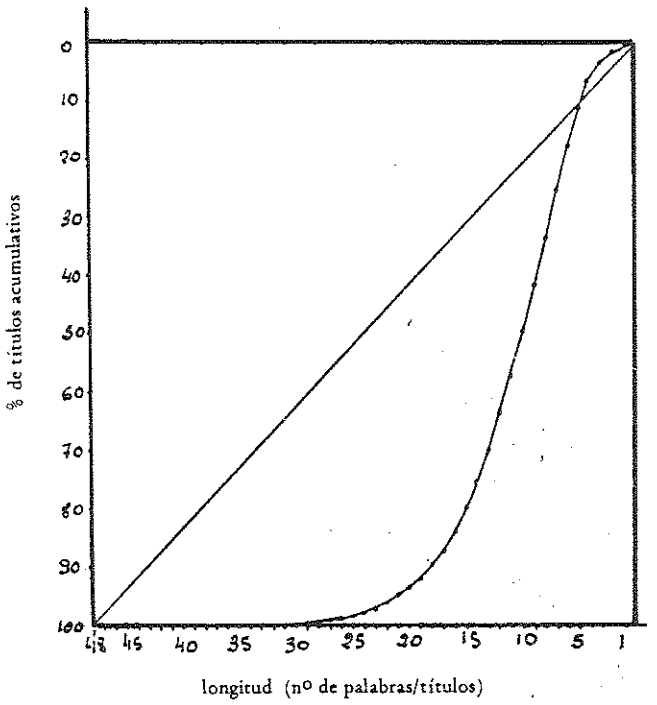


Figura nº3



En la Figura 2 se muestra la distribución estadística de las longitudes de los títulos, expresadas en función del número de palabras que contienen.

La Figura 3 muestra una curva de Lorenz que informa sobre el peso que tienen en la colección los títulos con un número de palabras inferior o superior a uno dado. De ella deducimos que:

– El 97,07 por ciento de los títulos de la colección contiene un número de palabras comprendido entre 3 y 20.

– El 6,29 por ciento de los títulos contiene menos de 5 palabras.

– El 6,37 por ciento de los títulos contiene más de 20 palabras.

– Los títulos con más de 32 palabras suponen un 0,30 por ciento del total.

Palabras

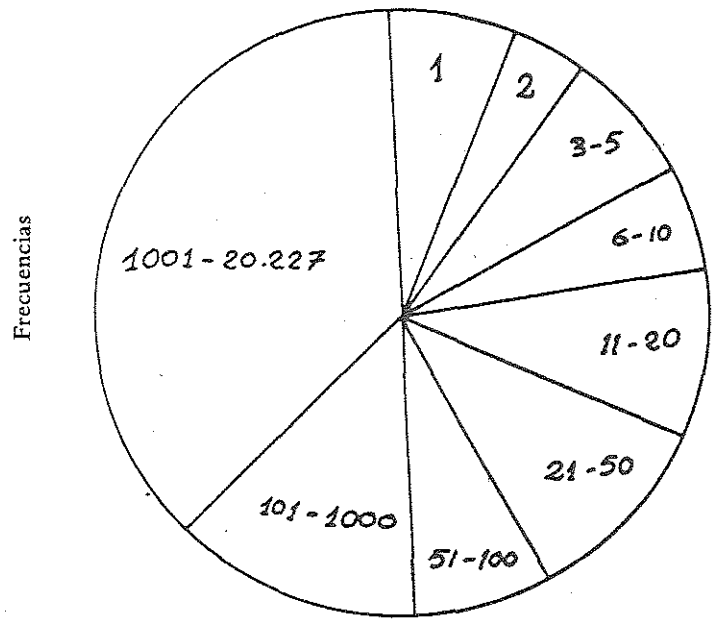
De las 142.944 palabras extraídas a partir de la colección de títulos, sólo 18.099 palabras eran diferentes, lo que supone un 12,66 por ciento del total. Dichas palabras muestran una frecuencia de aparición muy diversa, como puede deducirse de los datos expuestos en la Tabla 1; por consiguiente, también es muy variable la aportación de cada una de ellas a la cantidad total de palabras extraídas, según puede apreciarse en la Figura 4.

Tabla 1

PALABRAS

Frecuencia	Nº de palabras diferentes	% Diferentes
1	9853	54'44
2	2753	15'21
3-5	2709	14'97
6-10	1297	7'17
11-20	755	4'17
21-50	481	2'66
51-100	147	0'81
101-1000	91	0'50
1001-20227	13	0'07
Total	18099	100'00

Figura nº 4



Aportación de las diferentes palabras, considerando su frecuencias de aparición, al número total de palabras extraídas a partir de los títulos de la colección

Podemos comprobar que un 54,44 por ciento de las palabras diferentes aparecen solamente una vez en la colección; puede verificarse igualmente, a partir de lo expuesto en la Tabla 1, que las 2.784 palabras diferentes que poseen una frecuencia mayor de 5, y que supone el 15,38 por ciento de las palabras diferentes, contribuyen a la colección con un 82,27 por ciento del total de "ocurrencias" (frecuencia de aparición) en los títulos.

Asimismo, podemos advertir que 101 palabras de dicha lista –esto es un 0,56 por ciento de las incluidas en ella– aparecen en total 71.529 veces, lo que supone el 50,04 por ciento del total de ocurrencias. De entre esas 101 palabras más frecuentes hay 34 que consideramos "vacías", y de ellas destacan 20 (ver Tabla 2), que son precisamente las 20 palabras de mayor frecuencia de la lista, y que aparecen 57.425 veces, lo que supone el 40,17 por ciento del total de palabras contenidas en la colección de títulos.

En este trabajo se han considerado como palabras vacías a una serie de palabras funcionales, tales como artículos, pronombres, preposiciones, conjunciones, verbos auxiliares y otras palabras de bajo contenido semántico o de uso general y poco valor informativo; asimismo se han considerado como vacías algunas letras y números romanos de frecuente aparición.

Tabla 2

Rango	Frecuencia	Long.	Palabra
1	20227	2	DE
2	6781	2	LA
3	5693	2	EN
4	4407	1	Y
5	3353	3	DEL
6	2362	2	EL
7	2304	3	LOS
8	1755	3	LAS
9	1314	5	SOBRE
10	1263	4	PARA
11	1260	7	ESTUDIO
12	1142	1	A
13	1015	3	CON
14	944	3	POR
15	821	2	UN
16	611	1	I
17	607	2	AL
18	572	3	UNA
19	563	2	II
20	431	2	SU

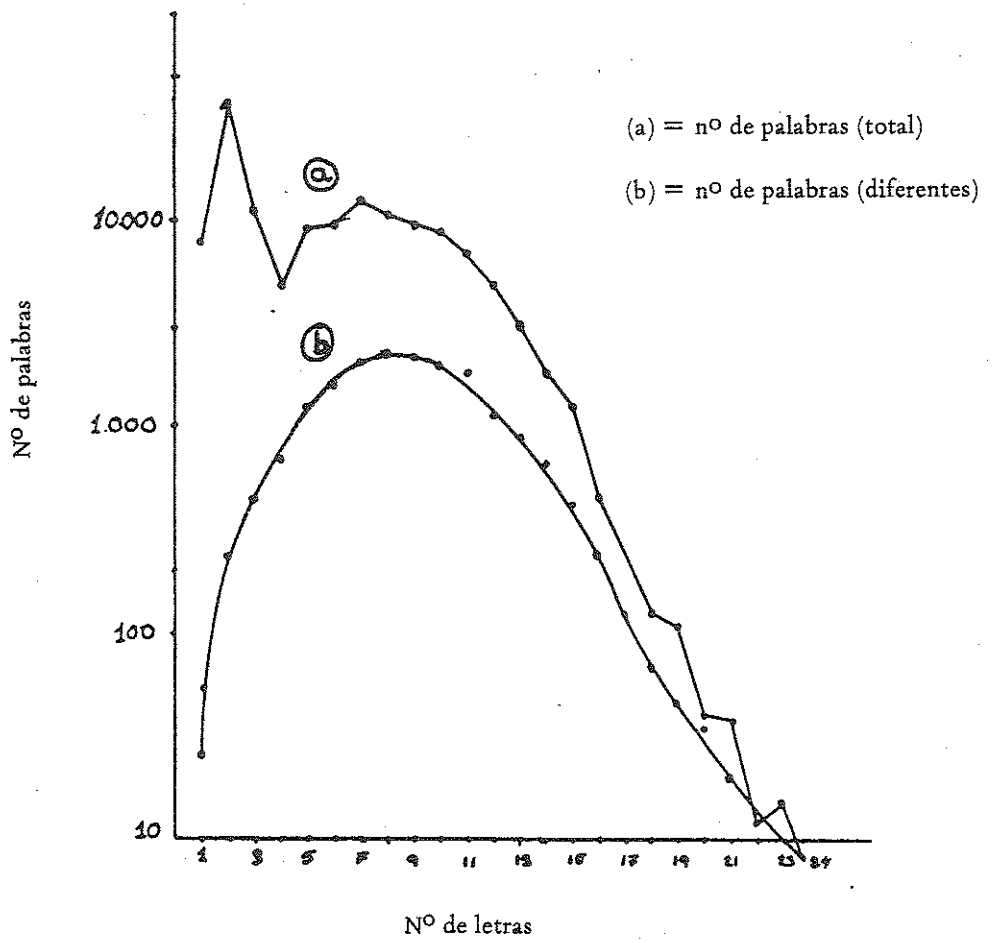
Caracteres

Considerando las 142.944 palabras extraídas a partir de los títulos de la colección, la longitud media por palabra ha resultado ser de 5,81 caracteres; pero si se tienen en cuenta sólo las 18.099 palabras identificadas como diferentes, la longitud media resultante es de 8,97 caracteres por palabra.

El número de palabras extraídas que tenían más de 24 caracteres ha sido de 18, de las cuales 16 eran diferentes. Todas ellas corresponden a nombres de compuestos químicos de estructura compleja, tales como CARBOXIETILENDIAMINOTETRAACETICO, DIAMINOCICLOHEXANOTETRACETICO, etc. Dada su escasa incidencia, y teniendo en cuenta que fijar en 32 el número de posibles caracteres de una palabra incrementaría desproporcionadamente la reserva de espacio necesaria para el almacenamiento de las palabras, se decidió excluirlas de la lista de palabras diferentes. Esta colección, ahora conteniendo 18.083 palabras, junto con los datos correspondientes a su frecuencia y longitud, han constituido la base de diversos estudios estadísticos de algunos de los cuales se detallan sus resultados a continuación.

La Figura 5 muestra dos curvas que representan: a) el número total de palabras extraídas, y b) el número de palabras diferentes que poseen una longitud dada comprendida entre 1 y 24 letras.

Figura nº 5



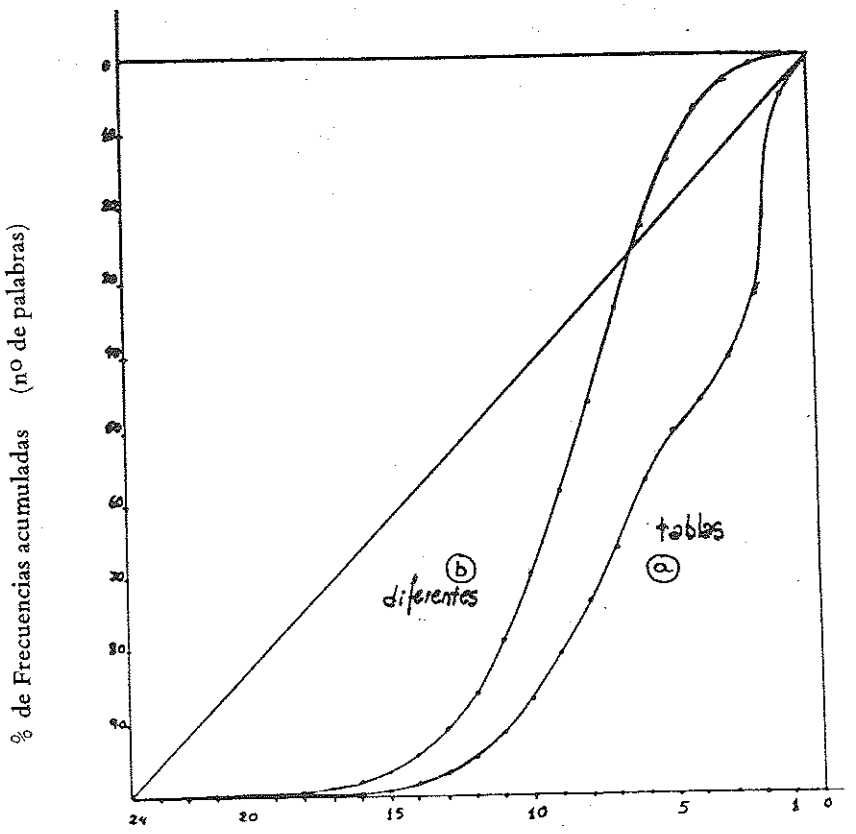
En la curva b) podemos apreciar cómo el número de palabras diferentes que tienen una longitud dada aumenta al hacerlo su longitud, en el caso de que ésta sea menor de 8 caracteres. En cambio, a partir de las 9 letras el número de palabras diferentes disminuye al aumentar su longitud.

En la curva a) vemos cómo el número total de palabras extraídas que poseen una longitud comprendida entre 1 y 3 letras inclusive no se corresponde con lo que cabría esperar de acuerdo con lo visto en la curva b). Esto es debido a que en esos grupos de palabras diferentes con longitudes de 1, 2, ó 3 caracteres

existen palabras que poseen frecuencias de aparición notablemente altas, como sucede por ejemplo con las palabras de una letra, Y, A, I; de dos letras: DE, LA, EN, EL, UN, II; o de tres letras: DEL, LOS, LAS, CON, POR, UNA.

Con los datos utilizados para representar las curvas a) y b) de la Figura 5 se han elaborado las correspondientes curvas de frecuencias acumuladas de palabras en relación con el número de caracteres por palabras que se muestra en la Figura 6.

Figura nº 6



En ella podemos comprobar cómo el número de palabras diferentes que tienen menos de 16 caracteres constituye el 96,86 por ciento de la lista de palabras diferentes; asimismo, podemos ver que el 99,28 por ciento de todas las palabras extraídas a partir de los títulos también tienen menos de 16 caracteres.

En la Tabla 3 se muestran las frecuencias de aparición de las letras simples que aparecen: a) en las 142.944 palabras extraídas de la colección de títulos; y b) en la lista de 18.099 palabras diferentes obtenidas a partir de dichos títulos.

Tabla 3

Ocurrencias de letras en las palabras contenidas en

(b) La lista de palabras diferentes			(a) Los títulos		
Rango	Letra	Frecuencia	Rango	Letra	Frecuencia
1	A	19576	1	E	99789
2	I	17165	2	A	95304
3	O	16431	3	O	79085
4	E	14405	4	I	77641
5	S	11887	5	S	60264
6	R	11501	6	N	56322
7	N	10190	7	R	50933
8	C	9951	8	C	50696
9	T	9222	9	L	49133
10	L	8181	10	D	48592
11	D	5916	11	T	40944
12	M	5165	12	U	23934
13	U	4815	13	M	21979
14	P	4027	14	P	19974
15	G	2463	15	G	9401
16	B	2378	16	B	9305
17	F	2271	17	F	8759
18	H	1709	18	V	7551
19	V	1571	19	Y	5565
20	Z	906	20	H	4087
21	X	703	21	Z	3033
22	Y	530	22	X	2146
23	J	419	23	Q	1754
24	Q	387	24	J	1585
25	K	220	25	Ñ	1351
26	Ñ	140	26	K	352
27	W	128	27	W	216

La diversa frecuencia de aparición de letras en títulos que refleja la Tabla 3ª es un reflejo de la desigual frecuencia de las palabras en los títulos de la colección.

Tanto la Tabla 3ª como la 3b ponen de manifiesto que son las vocales, con la excepción de la U, las letras que aparecen con mayor frecuencia, y que de entre las consonantes las más frecuentes son, por orden de más a menos,las letras S, R, N, C, T, L, D, M y P, siendo las menos frecuentes las letras J, Q, K, Ñ y W.

De la Tabla 3ª se deduce que son vocales el 45,23 por ciento de los caracteres que aparecen en las palabras contenidas en los títulos. De la Tabla 3b se deduce que las vocales consituyen el 44,68 por ciento de los caracteres presentes en los vocablos incluidos en la lista de palabras diferentes.

La Tabla 4 permite apreciar con qué frecuencia las diferentes letras simples aparecen como iniciales de las palabras extraídas: a) de la colección de títulos, o b) de la lista de palabras diferentes. A diferencia de lo visto en la Tabla

Tabla 4

1ª letra de las palabras (Frecuencias de aparición)

(b) Lista de palabras dif.			(a) Lista de palabras en títulos		
Nº de orden	Letra	Frecuencia	Nº de orden	Letra	Frecuencia
1	C	2155	1	D	28049
2	A	1814	2	E	16727
3	P	1680	3	L	13203
4	M	1280	4	C	12544
5	E	1210	5	A	10863
6	S	1132	6	P	10293
7	D	972	7	S	7149
8	T	950	8	M	6010
9	R	831	9	I	5206
10	F	829	10	Y	4508
11	I	724	11	T	3968
12	B	716	12	R	3922
13	H	668	13	F	3606
14	L	632	14	N	2726
15	G	574	15	V	2334
16	O	448	16	G	2238
17	V	435	17	U	2161
18	N	411	18	B	2059
19	U	172	19	O	1998
20	J	92	20	H	1896
21	K	81	21	Q	548
22	Q	68	22	Z	
23	Z	62	23	J	215
24	W	59	24	X	169
25	X	49	25	K	154
26	Y	39	26	W	104
27	Ñ	–	27	Ñ	–

3, los datos de la Tabla 4 no reflejan un predominio de las vocales sobre las consonantes, pues, como podemos deducir de la Tabla 4ª, las vocales representan

el 25,85 por ciento de las letras iniciales de todas las palabras (142.944) extraídas a partir de la colección de títulos. De la Tabla 4b deducimos que las vocales representan el 24,12 por ciento de las iniciales correspondientes a las 18.099 palabras diferentes.

Puede comprobarse asimismo que las letras que en la Tabla 3 aparecían como de escasa frecuencia aparecen también en la Tabla 4 como iniciales de escasa frecuencia. Finalmente, puede apreciarse en la Tabla 4b la notable incidencia que tiene sobre la frecuencia de las letras iniciales la presencia de palabras de alta frecuencia, tales como las incluidas en la Tabla 3.

Aplicaciones

El diccionario de frecuencias y los demás resultados obtenidos en este estudio constituirán la base de diversos estudios, en fase de realización, relacionados con la comprensión de datos, el almacenamiento y recuperación de información, y la detección y corrección de errores ortográficos y tipográficos.